[CANCER RESEARCH 63, 655–657, February 1, 2003]

Computational Analysis and Experimental Validation of Tumor-associated Alternative RNA Splicing in Human Cancer

Zhining Wang, H. Shuen Lo, Howard Yang, Sheryl Gere, Ying Hu, Kenneth H. Buetow, and Maxwell P. Lee[1]

Laboratory of Population Genetics, National Cancer Institute, Gaithersburg, Maryland 20877

ABSTRACT

A genome-wide computational screen was performed to identify tumor-associated alternative RNA splicing isoforms. A BLAST algorithm was used to compare 11,014 genes from RefSeq with 3,471,822 human expressed sequence tag sequences. The screen identified 26,258 alternative splicing isoforms of which 845 were significantly associated with human cancer, and 54 were specifically associated with liver cancer. Furthermore, canonical GT-AG splice junctions were used significantly less frequently in the alternative splicing isoforms in tumors. Reverse transcription-PCR experiments confirmed association of the alternative splicing isoforms with tumors. These results suggest that alternative splicing may have potential as a diagnostic marker for cancer.

INTRODUCTION

Most mammalian genes consist of multiple exons interspersed with introns. Introns are removed from a primary transcript by RNA splicing, which generates a mature translatable mRNA (1). In some cases, more than one mRNA is generated from a single primary transcript via an alternative splicing mechanism. It is well documented that alternative RNA splicing plays a biologically important function. For example, membrane-bound IgM and secreted IgD are produced by alternative splicing of the same primary transcript (2), and sex determination is regulated by alternative RNA splicing of *sxl*, *tra*, and *dsx* (3) in *Drosophila*. Several studies have analyzed alternative RNA splicing on a genome-wide basis (4–8). These studies showed that 35–59% of human genes have at least one alternative splicing isoform. Several studies show that alternative RNA splicing occurs frequently in human cancer cells (9–11). This study examines alternative splicing associated with cancer on a genome-wide basis and provides experimental validation for the tumor-associated alternative splicing. The results suggest that alternative splicing may have potential as a diagnostic marker for cancer.

MATERIALS AND METHODS

Computational Analysis of Alternative RNA Splicing. Each human RefSeq (NM_xxxxxx) sequence was compared with the human EST[2] database to identify alternative splicing isoforms using a BLAST parser. The algorithm defined an alternative splicing event as when two perfectly aligned regions between the EST and RefSeq were separated by a gap. An E-value of less than 10^{-10} was required for the flanking aligned sequences, and the gap was required to be >10 bp. The following z statistic was used to estimate the probability that an alternative splicing isoform is associated with tumor cells: $z = (p_t - p_n)/\sqrt{p(1 - p)(1/C_t + 1/C_n)}$. For a given alternative splicing isoform, p_t and p_n are the frequency of that alternative splicing event in tumor and normal libraries, respectively, and p is the average frequency of the alternative splicing event. C_t and C_n are the number of ESTs in the tumor and normal libraries, respectively. We used one-side z test for each alternative splicing site using the z-statistics defined above. Because the two proportions p_t and p_n are approximately normally distributed by the central limit theorem, the difference $p_t - p_n$ is also approximately normally distributed. Under the null hypothesis $p_t = p_n = p$, the variance is $p(1 - p)(1/C_t + 1/C_n)$. We defined the statistics z as the difference of the proportions divided by SD. So z is an approximately standard normal variable. An alternative splicing isoform is considered tumor-associated if it has a $P < 0.05$. Tumor-associated alternative splicing isoforms for the tissue-specific type of cancer were identified by selecting ESTs from cancer and normal tissues.

Experimental Validation. Primers were designed with Primer3 software (12) to detect regular and alternative splicing products. Primer sequences are available online.[3] Total RNA was isolated using RNAzol B (Tel-Test, Inc., Friendswood, TX) according to the manufacturer's protocol. cDNA synthesis was carried out using avian myeloblastosis virus reverse transcriptase (Invitrogen Corp., Carlsbad, CA) and oligo(dT)$_{12-18}$ primers (Invitrogen Corp.). A Packard MultiPROBE II EX robotic liquid handling system (Packard Instrument Company, Meriden, CT) was used for PCR. PCR was carried out in a 15-μl reaction in 1× buffer [1.5 mM Mg^{2+}, 0.2 mM dNTP, 0.5 μM primers, 1 unit of Taq DNA polymerase (Applied Biosystems)] and 2 μl of cDNA. Amplification parameters were as follows: 95°C for 10 min; 40 cycles of 95°C for 45 s; 60°C for 30 s; and 72°C for 60 s, followed by extension at 72°C for 10 min. Reaction products were analyzed by agarose gel electrophoresis. Reactions containing multiple PCR products were purified using the QIAquick purification kit (Qiagen, Inc., Valencia, CA).

To confirm the identity of PCR products, 18 PCR fragments were sequenced using ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems) and an ABI Prism 3100 or 3700 Genetic Analyzer (Applied Biosystems). Sequencing traces were analyzed using Sequencher software (Gene Code Corporation, Ann Arbor, MI) and Phred/Phrap (13). The results confirmed the expected DNA sequence in all cases.

RESULTS AND DISCUSSION

Genome-wide Screening for Tumor-associated Alternative Splicing Isoforms. Alternative RNA splicing isoforms were identified for 11,014 RefSeq sequences using a BLAST search algorithm to compare each RefSeq pairwise with 3,471,822 sequences in the human EST database. The algorithm identified 26,258 alternative splicing isoforms, one-third of which had two or more ESTs. The majority of ESTs aligned perfectly with a cognate RefSeq sequence (*i.e.*, regular splicing event); however, in many comparisons, two perfectly aligned regions were separated by a gap (*i.e.,* alternative splicing event). An EST that does not match its cognate RefSeq perfectly is an alternative splicing isoform. Each alternative RNA splicing isoform has unique parameters, including the location of the splice junctions and the length of the inserted or deleted sequence.

If alternative RNA splicing isoforms are associated with specific cancer cell types, then they could potentially serve as diagnostic markers for cancer. This idea was tested as follows. Alternative splicing isoforms were classified as tumor or normal based on the source of the mRNA used to construct the relevant cDNA library and *P* that the isoform was present at higher frequency in tumor cDNA libraries than in normal cDNA libraries (see "Materials and Methods" for details). This analysis showed that 845 alternative RNA splicing isoforms (3.2%) were significantly ($P < 0.05$) associated with tumor libraries. A complete list of 845 tumor-associated alternative splicing isoforms is available online.[3] The alternative splicing isoforms that

Received 9/12/02; accepted 12/2/02.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] To whom requests for reprints should be addressed, at Phone: (301) 435-1536; Fax: (301) 402-9325; E-mail: leemax@mail.nih.gov.

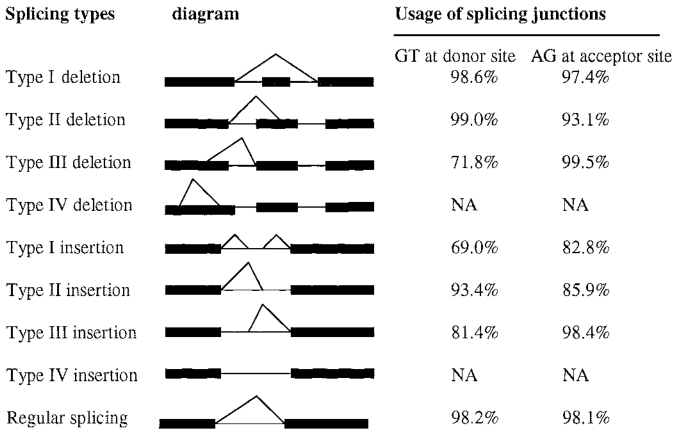
[2] The abbreviations used are: EST, expressed sequence tag; RT-PCR, reverse transcription-PCR.

[3] Internet address: leelab.nci.nih.gov/ASCA.

Table 1 *Identification of tumor-associated alternative splicing isoforms in various tissues*

Tissue	No. of ESTs		No. of significant isoforms[a]	Examples
	Tumor	Normal		
Liver	17,677	41,336	54	CYP2C8, KNG, AMBP, AHSG, ANG, TDO2, BHMT2, ADH4, HFL1
Brain	72,910	56,778	29	MAD2L1, RAD1, MAPK8IP1, PCMT1, MCK, POLR2I, PKIA, DLEU2
Placenta	24,515	72,100	26	LDHC, ADM, ATP6S1, ENG, IFNGR2, GPS1, DSPG3, HHLA2, UROD
Lung	55,083	27,968	6	LRP1, SFTPC, TCEB1L, SKP1A, RCL, HCG1V.9
Kidney	32,020	20,503	5	CD9, PCK2, HIBCH, TINAG, TSPAN-1
Prostate	17,667	17,225	4	TGM4, KLK2, PPP2R5A, HT012
All tissues	766,090	592,165	845	RAB1A, NME1, NEU2, TOP1, MRE11A, NCOA1, RNPEP, IRAK1

[a] This is the number of alternative splicing isoforms that are significantly associated with tumors in the given tissue; see "Materials and Methods" for details.



Splicing types	diagram	Usage of splicing junctions: GT at donor site	AG at acceptor site
Type I deletion		98.6%	97.4%
Type II deletion		99.0%	93.1%
Type III deletion		71.8%	99.5%
Type IV deletion		NA	NA
Type I insertion		69.0%	82.8%
Type II insertion		93.4%	85.9%
Type III insertion		81.4%	98.4%
Type IV insertion		NA	NA
Regular splicing		98.2%	98.1%

Fig. 1. Splicing junctions in eight alternative splicing subgroups. The splicing junction sequences were analyzed as described in "Materials and Methods" and "Results and Discussion." The percentage of GT-AG splice junctions are indicated for the junction unique to the alternative transcript. NA indicates that splice junction sequences are not known.

are associated with liver, brain, placenta, lung, kidney, and prostate cancers were also identified (Table 1). Some examples of tumor associated alternative splicing isoforms in various tissues were listed in Table 1.

Splice junctions in the alternative RNA splicing isoforms were analyzed and categorized into eight subgroups. The frequency was calculated for each splicing subgroup (Fig. 1). Interestingly, GT-AG splice junctions were used significantly less frequently in alternative splicing isoforms than in regular splice events. GC-AG splice junctions were used in 16.8% of type I insertions in alternative splicing isoforms but in ~0.43% of regular splicing events (39-fold increase). GC-AG usage increased specifically in tumors ($P < 0.002$ in χ^2 test).

Validation of Tumor-associated Alternative Splicing Isoforms by RT-PCR. To validate the computation-predicted alternative RNA splicing isoforms, 12 pairs of matched tumor and normal RNA samples from lung, breast, liver, and prostate tumors were analyzed by RT-PCR. Primers were designed to specifically amplify regular or alternative splicing products (Fig. 2). Seventy-six alternative RNA splicing isoforms were selected for analysis based on a $P < 0.05$, or because they were known to be involved in cancer. Fifty-five (72%) of the expected products were detected by PCR. Forty-five of these alternative RNA splicing isoforms were expressed in tumor but not matched normal samples. For example, three lung cancer samples expressed alternative splicing products, but none of their matched normal samples expressed the same mRNA isoform (Fig. 2, *Lanes 1–6*, the alternative splicing marked with deletion). However, this pattern was not observed for all samples; alternative splicing products were detected in tumor and matched normal samples from liver (Fig. 2, *Lanes 7–8*).

Alternative splicing is differentially associated with specific tissues or types of cancer. Some tumor-specific splicing isoforms were detected in several tumor tissues, whereas others were less prevalent (Fig. 3). For example, the NME1 alternative RNA splicing isoform was detected in all five lung cancer samples but not in any of the breast or liver cancer samples. However, the RAB1A alternative RNA splicing isoform was present in all four types of cancer tissues. Interestingly, three genes, RAB1A, RBBP8, and AXL, expressed alternative but not regular splicing variants in tumors. Conversely, the matched normal tissues expressed only regular splicing products. The absence of regular splicing products in tumors suggests that loss of regular splicing isoforms may be associated with tumorigenesis.

Our validation results show that 45 of 76 (59.2%) selected alternative splicing isoforms were tumor specific. This observation is consistent with previous reports that alternative splicing often increases in

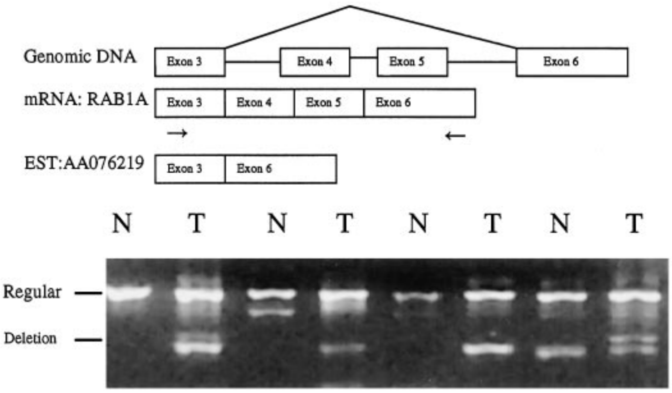


Fig. 2. Experimental validation of alternative splicing using RT-PCR. N and T denote normal and tumor, respectively. Regular and deleted splicing products are indicated. *Arrows* represent primers and *open boxes* designate exons. The *horizontal lines* between exons represent introns. *Diagonal lines* indicate alternative splicing events.

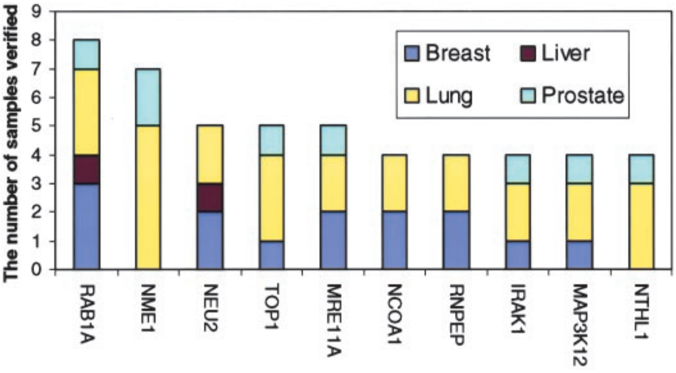


Fig. 3. Summary of tumor-associated alternative splicing validation experiments. Four breast, one liver, five lung, and two prostate cancer samples and their matched normal samples were used in the validation experiment. The Y axis represents the number of sample pairs that specifically expressed alternative splicing isoforms in the tumor sample and expressed regular splicing products in the normal sample. Regular splicing products may or may not have been detected in the tumors. Data are shown from 10 representative genes. Additional data are available online.[3]

human cancer (9–11), although the cause of the increase is not known. One possibility is that splicing fidelity is compromised in tumors because of mutations or altered gene expression affecting the splicing machinery. The mechanisms stimulating tumor-associated alternative splicing may be related to microsatellite instability in cancer cells, which is caused by defects in mismatch repair (14). Future experiments are required to test this hypothesis.

ACKNOWLEDGMENTS

We thank Drs. David Kaufman, Kent Hunter, and Tzu-ling Tseng for helpful comments on the manuscript.

REFERENCES

1. Hastings, M. L., and Krainer, A. R. Pre-mRNA splicing in the new millennium. Curr. Opin. Cell Biol., *13*: 302–309, 2001.
2. Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell, *20:* 313–319, 1980.
3. Nagoshi, R. N., McKeown, M., Burtis, K. C., Belote, J. M., and Baker, B. S. The control of alternative splicing at genes regulating sexual differentiation in D, melanogaster. Cell, *53:* 229–236, 1988.
4. Huang, Y. H., Chen, Y. T., Lai, J. J., Yang, S. T., and Yang, U. C. PALS db: putative alternative splicing database. Nucleic Acids Res., *30:* 186–190, 2002.
5. Modrek, B., Resch, A., Grasso, C., and Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res., *29:* 2850–2859, 2001.
6. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett., *474:* 83–86, 2000.
7. Mironov, A. A., Fickett, J. W., and Gelfand, M. S. Frequent alternative splicing of human genes. Genome Res., *9*: 1288–1293, 1999.
8. Xie, H., Zhu, W., Wasserman, A., Grebinskiy, V., Olson, A., and Mintz, L. Computational analysis of alternative splicing using EST tissue information. Genomics, *80:* 326–330, 2002.
9. Stimpfl, M., Tong, D., Fasching, B., Schuster, E., Obermair, A., Leodolter, S., and Zeillinger, R. Vascular endothelial growth factor splice variants and their prognostic value in breast and ovarian cancer. Clin. Cancer Res., *8:* 2253–2259, 2002.
10. Adams, M., Jones, J. L., Walker, R. A., Pringle, J. H., and Bell, S. C. Changes in tenascin-C isoform expression in invasive and preinvasive breast disease. Cancer Res., *62:* 3289–3297, 2002.
11. Lee, M. P., and Feinberg, A. P. Aberrant splicing but not mutations of TSG101 in human breast cancer. Cancer Res., *57:* 3131–3134, 1997.
12. Rozen, S., and Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol., *132:* 365–386, 2000.
13. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res., *8:* 175–185, 1998.
14. Kinzler, K. W., and Vogelstein, B. Lessons from hereditary colorectal cancer. Cell, *87:* 159–170, 1996.